Open your eyes and listen carefully. Auditory and audiovisual speech perception and the McGurk effect in aphasia

Klitsch, Julia Ulrike

2 AUDITORY AND AUDIOVISUAL SPEECH PERCEPTION

Prior to approaching questions characteristic of the speech perception process, the speech perception process itself should be defined, immediately posing the first problem. In the strict sense, speech perception is an acoustic-phonetic process that starts with the acoustic signal and at the end of the process stands a basic phonological unit, with meaning being traditionally ignored. However, there is the spoken word recognition process, which starts with the output from the speech perception process, that is, the basic phonological unit and the output is a (lexical) word. This subprocess ignores phonetic details. However, the human processing system does not appear to be such a strictly segregated system, but processes and stages appear to overlap, to a large extent. This is reflected, for example, by findings like the Ganong effect (Ganong, 1980), in which phoneme boundaries are shifted according to lexical status. That is, an ambiguous phoneme, that could be, for example, either /t/ or /d/, is more likely to be perceived as /t/ in the context **ype*, since *type* is a word whereas *dype* is not. On the other hand, the target phoneme /d/ would be the more likely percept in a context like **ash* because /d/ gives rise to a word, *dash*, whereas /t/ *tash* does not. Other phenomena that illustrate the close entanglement of purely perceptual and lexical-semantic processes are the word superiority effect, or the perception of speech as a bottom-up or top-down process (2.4). Since this study deals with the human processing system, speech perception should be seen in this broader sense, instead of the strictly phonetic sense.

The daily encounter with speech – a continuous, seemingly boundary-free flow of sounds that reach our ears in face-to-face communications, conversations via the telephone, or when listening to the radio – seems to pose a challenge but it is usually mastered effortlessly. Effortlessly, at least, when speaker and listener have the same language-background or when there is not too much distorting noise in the listening condition, be it environmental noise or noise in the metaphoric sense in the perceiver, caused, for example, by aphasia.

How does the listener approach spoken speech? How does he or she know what are the relevant cues in the auditory or audiovisual signal? How do pieces of information from audition and vision interact in the perception of speech and how and at what point are they integrated?

This chapter describes the auditory (2.1) and audiovisual speech perception process (2.2), accounts for speech perception in terms of bottom-up and top-down processing (2.3), and illustrates speech perception impairments (as a possible cause of speech comprehension deficits) in aphasic patients (2.4).

2.1 Auditory speech perception

Auditory speech perception is a complex process that centers on a quickly fading incoming auditory speech signal that is lost as fast as 400 ms if the information comprised in the speech signal is not immediately converted into a coherent sequence of discrete speech units (e.g. Pisoni & Tash, 1974). This ephemerality of auditory speech properties, however, challenges a purely auditory account that assumes speech perception to be determined by the success of fitting an encountered, yet unrecognized sensory form into an appropriate, memorized auditory template. Given the perishability of the raw auditory speech properties, and the continuous speech stream with no direct match between acoustic information and associated memorized phonemes, this seems a clear limitation for a purely auditory account of speech perception. Pardo and Remez (2006) suggest that the perishable auditory properties of speech may still access the listener's memory by adopting another form. Hirsh (1988) proposes the adopted form may be defined in terms of the source evoking the auditory sensations.

In addition, the fact that speech is not usually encountered in quiet listening conditions but in situations mixed with numerous other speech and non-speech sounds could be an additional hindrance. This requires the listener to extract and integrate the linguistically relevant acoustic information from an auditory signal heaped with relevant and irrelevant information. Opinions diverge concerning the nature or specificity of this information filter. Sometimes it is assumed that the processor that extracts linguistic information from the incoming periodic and aperiodic sequences of noise that are intermitted with periods of silence is just an unspecific acoustic processor that deals with any kind of sound, be it speech or not. Alternatively, a phonetic-phonological processor as part of the human-specific genetic endowment is assumed. This domain-specific processor is sometimes seen as a domain-specific module that is supposed to be tuned to the fast, automatic, and unconscious processing of specific structures. In addition, if there was indeed such a module it would be assumed to be operationally encapsulated (Fodor, 1983). This means this module or processor is supposed to operate quickly and automatically at the encountered continuous, non-linear, invariant, information-loaded auditory signal. That

is, this module is believed to unzip the utmost complex, fast incoming auditory speech signal into linguistically relevant information (at conversational speech rate: up to 15 phonemes per sec or seven syllables per sec; Pollack & Pickett, 1964) and to operate on nothing else but speech structures.

Non-linearity means that there is no one-to-one-match between the acoustic signal and the perceived linguistic segments, such as phonemes, syllables, or words. Coarticulation, for example, causes an overlap between the acoustic properties of (two) abutting segments in the speech signal. The lack of clear boundaries between segments and the reciprocal influence of adjacent segments in turn makes it difficult to segment the continuous speech signal into the perceived linguistic units.

Invariance concerns the phenomenon that there are (combinations of) acoustic properties that are considered to be characteristic for a particular phoneme. However, the acoustic specification seems, crucially, to depend on the phonetic context in which this particular phoneme occurs and is, thus, subject to context-specific changes or adaptations. That is, the alleged invariant acoustic properties characterizing a particular phoneme are not exactly invariant across contexts, but seem to be highly context-sensitive. This lack of invariance requires the listener to match cues of differing acoustic qualities with a single phonemic concept. Other factors that spoil the idea of invariantly reliable cues to phonetic features are different speakers, differences in speech rate, and emotional conditions. For example, in more colloquial contexts, one and the same speaker articulates less carefully than in more formal contexts. Also, the configuration of the vocal tract as determined amongst others by gender or age of the speaker causes variation in the physical nature of the speech signal.

There are at least two phenomena that caused some researchers (e.g. Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967) to assume that the perception of speech is special: duplex perception and categorical perception. Hence, speech perception was attributed to the accomplishment of the aforementioned speech or phonological processor.

Duplex perception describes the fact that parts of the acoustic signal bear information relevant for both speech and non-speech perception. Several experiments have been performed in which the third formant transition was isolated from a continuum ranging from /ba/ to /ga/ (e.g. Ciocca & Bregman, 1989; Vorperian, Ochs, and Grantham, 1995). The third formant transition distinguishes between the two stimuli, whereas the base is identical for the stimuli. Presenting the base to one ear and an isolated third formant transition to the other, the identity of the perceived syllable depends on the kind of transition. When a formant transition is presented in isolation, or when asked to focus on

the non-speech portion of duplex stimulus, the formant transitions are perceived as non-categorizable, non-speech chirps or whistles. That is, the formant transitions (or isolated chirps) are used in speech perception, but speech perception cannot be solely explained by the processing of the non-speech chirps. Hence, the phenomenon of duplex perception has been interpreted as to indicate the presence of two autonomously operating systems: a general auditory and a phonetic-phonological perceptual system.

The finding that speech was perceived categorically caused some researchers (e.g. Liberman et al., 1967) to assume that speech perception is special. Liberman, Harris, Hoffman, and Griffith (1957) were the first to report categorical perception of speech with an auditory discrimination task. Two different types of auditory discrimination tasks are most often used. The first one follows the AX-paradigm, meaning that a stimulus pair is presented, with the two items either being identical (X is identical to A) or not (X differs from A). The task of the participant is to indicate whether the two items constituting a pair are identical or not. That is, the subjects have to react by replying either 'yes' or 'no' to a question, such as, 'Are the two stimuli identical?' The other variant makes use of the ABX-paradigm. In this case, two different template stimuli A and B are presented, directly followed by the test item X, which is identical to either A or B, and subjects have to judge whether X is identical to A or B. Liberman and colleagues (1957) found that subjects readily differentiated between two stimuli when they come from different categories (/ba/ vs /pa/) but not when they belong to the same category (e.g. two different versions of /ba/). This, despite the fact that all stimuli (intra- and intercategorical) were taken in equal steps from along the continuum; that is, all stimuli were, when physically considered, equally different. The authors assumed that categorical perception was typical for speech and concluded that speech perception was closely related to the speech production process (as reflected in the later postulated motor theory of speech perception; see 3.2.3.2). This assumption was based on the fact that a speaker either produces one phoneme or the other (for example, a voiced phoneme, or its voiceless counterpart). Consequently, they supposed that the listener simply perceives the same categories that he or she produces and nothing in-between.

With another task, an auditory identification task, Warren (1970) shows that at a certain point along the continuum there is an abrupt change in identification. Categorical perception, thus, means the failure to discriminate within speech categories and the presence of a sharp, distinct switch during identification.

However, there are a number of challenges to the assumption that speech perception is categorical. For example, if categorical speech perception is considered to be the result of

a phonetic-phonological (rather than an acoustic) processor, the findings of Kuhl (1987) that chinchillas also show categorical speech perception along a voicing continuum in the absence of such a speech-specific processor are difficult to account for. Also, if categorical perception were a characteristic of speech perception, it should apply to all speech sounds alike. There are, however, studies that failed to show categorical speech perception with vowels (e.g. Fry, Abramson, Eimas, & Liberman, 1962).

In addition, the mere fact that during speech perception discrete categories are perceived does not mean, per se, that speech is perceived categorically. For example, in studies in which speech stimuli from along a continuum had to be identified, longer reaction times were found for the stimuli that were closer to the category boundary – that is, for the less prototypical variants of a given phoneme (e.g. Studdert-Kennedy, Liberman, & Stevens, 1963). If speech sounds were indeed classified categorically, subjects should not be able to discriminate between sounds within one category as was suggested by the differences in reaction time (Massaro & Cohen, 1983). The results of a rating experiment by Massaro and Cohen (1983) are also interpreted as voting against categorical perception of speech. In this experiment subjects had to judge the degree of category membership of a given stimulus. The ratings suggested that subjects were aware of the continuous changes along a given (voicing) continuum, as reflected by relatively continuous rather than discrete perceptual changes.

In summary, the findings of these studies challenge the assumption that speech perception is categorical and favor the view of continuous speech perception, be it in auditory or multimodal (audiovisual) terms (Massaro, 2001).

Despite the lack of invariance and despite the non-linearity of the auditory speech signal, the listener readily and quickly processes the incoming speech stream that contains reams of overlapping acoustic cues. Some of these cues contained in the auditory signal are static, as are, for example, formants. In addition, the auditory speech signal also contains dynamic cues – formant transitions, for example – that reflect changes in the course of energy concentration at the crossing between consonant and vowel. Both types of acoustic cues have been considered to carry most weight in the perception of speech. Massaro (2001) annotates that during speech perception the listener probably draws on many different (all available) cues simultaneously and isolation of a single cue is assumed to be rather ineffective. Hence, it is assumed that the saliency of different single cues may vary from context to context and that, therefore, the combination and integration of all available cues (not only auditory, but also visual cues, for example) most reliably identify speech segments in the speech signal.

In the next section, the focus is on audiovisual speech perception. In many everyday language contexts (consider, for example, telephone conversations or listening to the radio), speech appears to be predominantly auditorily submitted. However, there are also findings that show that speech perception can be influenced and/or modified by visual articulatory information.

2.2 Audiovisual speech perception

The predominant view of speech perception as an auditory process is challenged by the fact that visual speech information has an effect on speech perception (e.g. Sumby & Pollack, 1954; Miller & Nicely, 1955; McGurk & MacDonald, 1976), suggesting that speech perception is perhaps more accurately described as a multimodal process, somehow gauging the auditory and visual speech input.

As with unimodal speech perception, the understanding of multimodal speech perception still contains a number of unresolved, controversial issues. There are two main questions: namely, whether the auditory and visual speech information is initially processed separately, in parallel, and only integrated at a later processing stage, or whether they are amalgamated right away and subsequently processed as a single piece of (multimodal) information. The second question concerns the integration frame. If the information streams are in the first instance being processed in parallel and integration occurs only after the phonemic form has been determined, then it can be assumed that the visual information is integrated into the phonological frame based on auditory information.

However, if intersensory blending occurs early, a common integration frame has to be identified. To identify this common integration frame seems to be quite a challenge since the sensory characteristics of the two modalities are not easy to group together due to their dissimilitude. Pardo and Remez (2006) report a study by Rosen, Fourcin, and Moore (1981) that provides support for the assumption of a common integration dimension of multimodal speech. In this study subjects were presented with a video of the face of a speaker, combined with the pulsing of the same speaker's larynx as amplified by means of an electroglottograph signal. The auditory component was rather awkward and not at all intelligible. There was nothing peculiar about the visual information, and even though visual speech perception exceeded auditory speech perception, it was poor as well. Contrary to the expectations that the integration of two insufficiently informative modalities will not yield a clear percept, the perception of the combined audiovisual speech was good. According to Pardo and Remez (2006) this result reflects the efficiency of integrating

audiovisual information and this integration into a common space presumably takes place before separate analyses of the information in the respective modality. However, given the large and, at first glance, seemingly incompatible differences in the sensory quality of the auditory and visual input, plus the reported tolerated temporal and spatial mismatches between the auditory and the visual input signal, it is rather striking how readily an audiovisual speech signal is integrated by the listener. In 3.2, models of audiovisual speech perception are described which deal with the non-simultaneous nature of the auditory and visual inputs as well as their relative collaborative effects on the final percept.

2.3 Speech perception as a top-down versus a bottom-up process

Another topic in speech perception research is whether speech is being processed bottom-up or top-down; that is, whether speech processing proceeds from a low to a higher level or vice versa. This discussion also illustrates the previously mentioned amalgamation of the (theoretically) strictly segregated processes of speech perception and spoken word recognition during natural speech perception.

Bottom-up processing starts at a low peripheral level and is passed through to higher, more central levels. In this case, the acoustic speech signal would be initially segmented into phonetic or distinctive features that, in turn, are assembled phonemes of which, finally, syllables and words are composed. This bottom-up 'assembly process' is not influenced by lexical context.

For *top-down* processing the opposite processing direction is assumed. High-level information, such as a listener's lexicon and contextual knowledge, generates expectations of (contextually) adequate input and, thus, can regulate lower-level processes. The almost infinite number and perishable nature of the acoustic pieces of information gave some researchers reason to deem top-down processing more feasible (e.g. Paran, 1997). This is because it is considered impossible that the human auditory system processes all this information separately and impartially.

Behavioral effects, such as the phoneme restoration effect, were interpreted in terms of providing evidence that speech processing can be influenced by the listener's knowledge and expectations and, hence, favoring top-down processing. The phoneme restoration effect was first shown by Warren (1970) and refined by Warren and Warren (1970). The authors show that when replacing a phoneme in a word with white noise or a cough and embedding it in a sentence context, the listener will understand the contextually most appropriate word without even being aware of the substituted phoneme. In their experiment they used the

sequence '*eel', with the star symbolizing the substituted phoneme. Depending on the sentence context, or more precisely the last word in the sentence, *eel was heard as 'wheel', 'heel', 'meal', or 'peel', respectively.

It was found that the *eel was on the axle.
It was found that the *eel was on the shoe.
It was found that the *eel was on the table.
It was found that the *eel was on the orange.

Findings that suggest that aphasic auditory speech comprehension deficits are related would be compatible with the view of speech processing as a bottom-up process (e.g. Luria, 1947; Tallal & Newcombe, 1978; see next section).

2.4 Speech perception in aphasia

Luria (1947) attributed the auditory speech comprehension impairment observed in patients with Wernicke's aphasia to deficits in what he called 'phonemic hearing' (cf. also Tallal & Newcombe, 1978). Even though there are studies that demonstrate that auditory discrimination is impaired to varying degrees in most aphasic patients (e.g. Blumstein, Baker, & Goodglass, 1977) the relation to auditory comprehension deficits is less clear. At first glance, the results by Blumstein and colleagues (1977) suggest a correlation between auditory speech discrimination and auditory speech identification abilities. However, the authors limit their own findings by adding that their results were mainly due to a group of patients with Broca's aphasia, while the other patients in their study (Wernicke's aphasics and patients with mixed anterior aphasia) did not show such a correlation. Rather, it seemed that in the group of Wernicke's patients this putative correlation was reversed; this group showed the most impaired language comprehension and the relative best scores on the phoneme discrimination task. In other studies, the lack of a correlation between auditory speech discrimination and auditory speech identification impairments was even more explicit. For example, Miceli, Gainotti, Caltagirone, and Masullo (1980) found that aphasic patients performing as clearly impaired on an auditory discrimination task can still show relatively preserved auditory speech comprehension and vice versa (also see Caplan & Aydelott-Utman, 1994). The opposite pattern, however, of better auditory analysis than auditory speech comprehension, occurred less frequently. Blumstein and colleagues (1977), for example, showed that the patients with Wernicke's aphasia were impaired at a phoneme

identification task, while their performance on a phoneme discrimination task fell in the normal range. In addition, all patients in this study (Wernicke's, Broca's, and mixed anterior patients) improved at a phoneme discrimination task for words compared to phoneme discrimination in non-word stimuli. This was interpreted to exclude a low-level impairment, which predicts that words and non-words are equally affected. Rather, the authors suggested that the results reflect an inability to assign a linguistically relevant label to an encountered acoustic stimulus. In any case, the data are not compatible with Luria's thesis that the auditory speech comprehension deficit of (Wernicke's) aphasics is attributable to impaired phonemic hearing.

Miceli and colleagues (1980) further annotated that other factors than pure auditory speech processing disorders (such as reduced attention span) may be involved in the decline of phoneme discrimination performance. However, this cannot fully account for the shortage of a correlation between phoneme discrimination and auditory language comprehension performance. The assumption that attentional factors can also contribute to the impaired phoneme discrimination performance in aphasic speakers was confirmed by Gow and Caplan (1996). In their studies a distinction was made between monosyllabic (CV vs. VC) and bisyllabic stimuli (VCV) and between different contrasting phonetic features (articulator-free [manner of articulation, sonorance] vs. articulator-bound [voicing, place of articulation]). Furthermore, they distinguished between between natural speech stimuli and synthesized speech stimuli. In addition, they used two different tasks: a phoneme discrimination trask and a phoneme identification task. Although performance of the aphasic patients was impaired at all controlled features, the patient group showed the same general performance pattern as the non-brain-damaged control group. This finding was true for various comparisons on the spoken phoneme discrimination task (consonant vs. vowel; voice vs. place of articulation; sonorant vs. non-sonorant contrasts; articulator-bound vs. articulator-free contrasts; CV vs. VC). In combination with a lack of significant interactions between subject type and feature type, the authors suggest that rather than indicating a specific deficit in acoustic-phonetic processing these findings could indicate a more general decline in performance due to more global factors, such as attention. Alternatively, Gow and Caplan (1996) suggest that phoneme discrimination performance is a measure of general processing complexity across populations. Thus, more complex processes could be more susceptible to aphasic disorders since they make greater demands on the processing system by requiring the integration of a greater amount of information, or by the mere fact that processing requires more neural tissue. The collected data are assumed to mirror the

relative vulnerability of distinguishable processes in a population with restricted processing resources (Gow & Caplan, 1996).